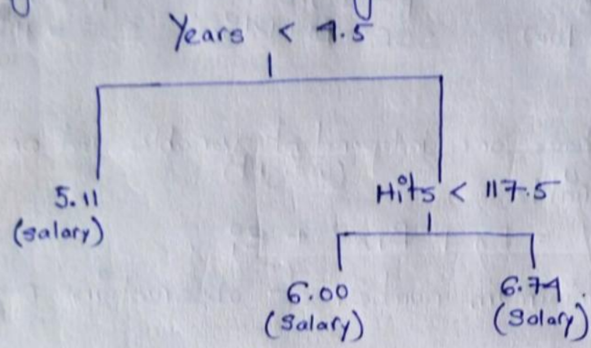
Decision tree case study - REGRESSION

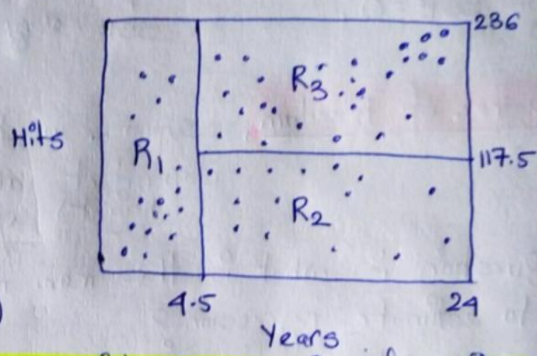
Predicting Baseball players Salaries Using Regression Trees.

Dataset → Salary, based on years (the number of years that he played in the major ~~years~~ leagues).

Hits, the numbers of hits that he made in previous years.

Fitting the data, tree is generated →





Overall, the tree stratifies or segment the players into three region of predictor space:

$R_1 = \{x \mid \text{Years} < 4.5\}$, $R_2 = \{x \mid \text{years} >= 4.5, \text{Hits} < 117.5\}$.

$R_3 = \{x \mid \text{years} >= 4.5, \text{Hits} >= 117.5\}$.

- In tree analogy, R_1, R_2 and R_3 are known as terminal node or leaves of tree
- The point along the tree where the predictor space is split are referred to as internal node. Two internal nodes are Years < 4.5 and Hits < 117.5

Story / Interpretation of tree →

- Years is the most important factor in determining Salary and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of hits that he made in previous years seems to play little role in his salary.
- But players who have five or more years of experience, the number of hits made in previous years does affect salary and players who made more hits last years tends to have high salaries.

Process of building a regression tree →

Mainly two steps

1) We divide the predictor space - that is, set of possible value for $X_1, X_2, X_3 \ldots, X_p$ - into J distinct and non-overlapping regions, $R_1, R_2, \ldots R_J$
2) For every observation that falls into region R_j, we make the same prediction, which is simply the mean of response values for training observation in R_j.

For instance, suppose that in step 1 we obtain Region R_1 and R_2.
The response mean of training observation in first region is 10.
Response mean of training observation in second region is 20.
Then for given observation, $X = x$ if $x \in R_1$ we will predict value of 10 and if $x \in R_2$ we will predict a value of 20.

How do we construct region $R_1, R_2, \dots R_J$?

→ The goal is find boxes $R_1, \dots, R_J$ that minimizes the RSS.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y - \hat{y}_{R_j})^2,$$ where $\hat{y}_{R_j}$ is mean response for training observation with j^{th} box.

y is actual value of set.

→ Unfortunately, computationally infeasible. That is why we use Recursive Binary Splitting.

Recursive Binary Splitting —

- Top-down, Greedy approach.
- Top down because it begins at the top of the tree (all points belong to single region)
- Then successively split the predictor space, each split is indicated via two branches further down.
- It is greedy because at each step of tree building process, best split is made at particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.
- We split the node, based on homogenity. We check the entropy and information gain.

Tree Pruning —

- Trees get overfit / underfit, leading to poor test set performance. This is because the resulting tree might be too complex / too simple.
- A smaller tree with fewer split might lead to lower variance.
- Larger tree may lead to high variance.

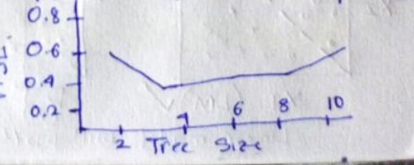
MSE can be used only for Regression based solution.

How do we determine the best way to prune trees?

- Goal of subtree that lead to lowest test error rate.
- We can use cross validation or validation set approach (but cost complexity)

Try to plot, MSE (Mean square Error) vs Tree Size.

As we can observe if we have 4 nodes, it have lowest MSE, then choose 4 node



SUM OF SQUARES

- Residual Sum of Square is used to help you decide if a statistical model is good fit for your data.
- It measures the overall difference between your data and the values predicted by your estimated model. (A residual is a measure of distance from a data point to a regression line).

Total Sum of Square = Explained sum of Square + Residual sum of Square.

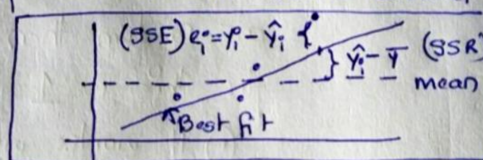
TSS = ESS + RSS.

$SSR = \Sigma(\hat{y} - \bar{y})^2 \quad SSE = \Sigma(y_i - \hat{y})^2$
$SST = SSR + SSE = \Sigma(y_i - \bar{y})^2$
$R^2 = SSR / SST$

→ TSS (SST) → Also known as TSS or SST.
- Tells how much variation is there in dependent variables.

$$TSS = \Sigma(y_i - \text{mean of } y)^2.$$

- SS (Sum of square) is a measure of how data set varies around a central number (like a mean).
- TSS is Σ (summation) of SS.



→ ESS (SSE) → Explained SS tells you how much of the variation in the dependent variable your model explained. ~~[From mean to the best fit is expected error and from best fit to point is unexpected / residual error]~~

$$\text{Explained SS} = \Sigma(\hat{y} - \text{mean of } y)^2.$$

→ RSS (SSR) → Residual SS tells you how much of the variation in the dependent variable you DID NOT EXPLAINED. (From mean to line was expected and from line to point is unexpected (residual))
- It is sum of squared difference between actual Y and predicted Y mean.
- The residual sum of square (RSS) also known as sum of square error (SSR) or sum of square estimated error (SSE) is the sum of square of residuals. (deviation predicted from actual set of values). $RSS = \Sigma(e^2)$.

NOTE - We normally do a square because of negative number. If the line is best fit, and we sum both negative and positive number then it will result to 0 (zero). So we square in order to avoid negative number. i.e, $\Sigma e = 0$
$\Sigma(\text{error}) = 0$.

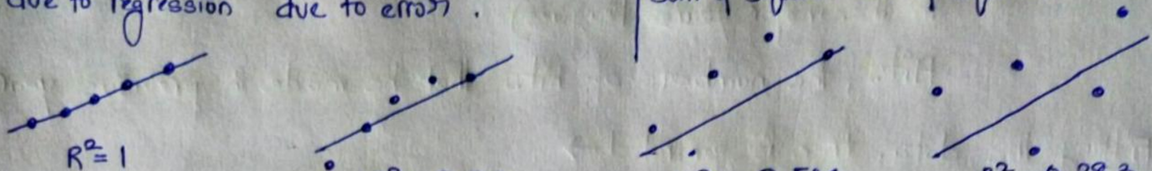
→ So we square $\Sigma(e^2)$, which is obviously greater than 0. i.e, $\Sigma(e^2) > 0$
And we try to minimise as much as possible $\min(\Sigma(e^2))$.

→ R^2 is proportion of total variation which is explained. High SSE, high error, low R^2. Low SSE, low error, High R^2.
Refer to page 6 (For figure)

R^2 = SSR / SST, The proportion of variation in Y being explained by variation in X.

SSR + SSE = SST | ESS + RSS = TSS.

Sum of square due to regression, Sum of square due to error. | Explained Sum of square, Residual Sum of square



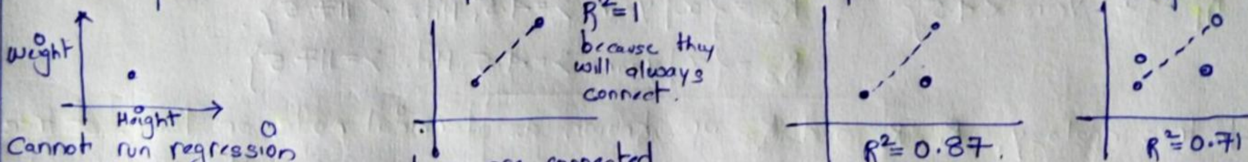
$R^2 = 1$, SSE = 0. | $R^2 = 0.83$, SSE = ~~very~~ low | $R^2 = 0.591$, SSE = ~~to~~ Medium | $R^2 = 0.283$, SSE = High

Degree of Freedom, Suppose we have one independent variable (weight) and one dependent variable (height).

$$Y = B_0 + B_1 x_i + \varepsilon_i.$$

Question is what is the ~~max~~ minimum number of observations required to estimate regression?

If we take one point | 2 points | 3 points | 4points



Cannot run regression | Lines are connected | $R^2 = 0.87$. Degree of freedom (DF) = 1 (points differ from model) | $R^2 = 0.71$ DF = 2.

Degree of Freedom = N − K − 1
(N: Total point; K: Independent Variable)

DF = 3 − 1 − 1 = 1. | DF = 4 − 1 − 1 = 2.

Observation, Increase number of Independent variable Degree of Freedom decreases.

If we have 2 independent vars, then DF = 3 − 2 − 1, DF = 0 | DF = 4 − 2 − 1, DF = 1.

Degree of freedom closely related to R^2.

As degree of freedom (df) decreases, R square will only increase.

In short more variables added to a given model, R^2 increases even if we give useless variable R^2 will increase.

Adjusted R^2, $R^2 = \begin{cases} 1 - (1 - R^2)\dfrac{n-1}{n-K-1} \\ \text{or} \\ 1 - \left(\dfrac{SSE}{SST}\right)\dfrac{n-1}{n-K-1} \end{cases}$

as K increases, Adj R^2 will tend to decrease, reflecting the reduced power in model. i.e, only add useful variable then only adj R^2 will increase